Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Análisis Predictivo aplicado al otorgamiento de préstamos hipotecarios para Dream House Finance

AUTOR: JUAN MANUEL ORQUIN

MENTOR: PABLO MATÍAS HERRERA

DICIEMBRE 2020

Resumen

En la era del Big Data, las empresas han iniciado un nuevo modelo y han desarrollado negocios de préstamos por medio de Internet, que sin duda presentan grandes desafíos al negocio tradicional de los bancos conservadores. Hoy en día, las empresas dadoras de préstamos hipotecarios exigen a los solicitantes la presentación de documentación personal, laboral y la del inmueble a adquirir. Este cúmulo de papelería es analizado, estudiado y trabajado por personal calificado de las entidades prestamistas generalmente de forma manual.
Para poder afrontar y adaptarse a la transformación digital y a la constante evolución de la tecnología, los bancos deben empezar a utilizar los grandes volúmenes de datos, provenientes de los sistemas de información, dispositivos electrónicos y redes sociales. Los datos están en manos de las entidades y pasaron a ser un activo valioso. Dicho de otra forma, los bancos pueden reducir los gastos estableciendo una estrategia de Big Data, innovando con un modelo predictivo para la toma de decisiones al momento del otorgamiento de préstamos.

Palabras claves: Big Data, préstamos hipotecarios, Análisis predictivo, Decision Tree, Naive Bayes

Índice

Introducción

En la actualidad la calificación para el otorgamiento de préstamos hipotecarios tiende a tener un gran porcentaje de tratamiento digital. Tanto la tendencia como el porcentaje van en aumento por lo que las entidades bancarias deben tratar de anticiparse a la competencia y a las demandas de los clientes. Las empresas deben adecuarse al cambio y a la transformación para continuar con el negocio. La generación y el análisis de datos aumentan su relevancia en el mundo en que vivimos. Las organizaciones tienen que empezar a utilizar técnicas de Big Data y Data Mining para procesar y analizar información valiosa en datos estructurados y no estructurados. Valerse de estas técnicas en forma efectiva impulsará la innovación en el modelo comercial y el modelo de toma de decisiones. Por lo que para lograr lo deben tener presente ¿Qué técnicas de minería de datos se deben utilizar para realizar una mejor clasificación binaria de los solicitantes a préstamos hipotecarios para determinar si es posible su pre-otorgamiento de manera virtual?

En este trabajo se desarrollará en el marco teórico las necesidades de la empresa Dream House Finance, dedicada al otorgamiento de préstamos hipotecarios, con el objetivo de lograr por medio del uso de Big Data y de la minería de datos una mejor clasificación binaria de los solicitantes. Para lograr este objetivo primero se trabajó sobre los aspectos organizacionales de la empresa, donde se definió la organización, cómo es la gestión de los datos, cómo es el proceso actual del análisis de otorgamiento y su problemática sobre la gestión de los datos. Posteriormente se abordó un set de datos de las solicitudes de préstamos previamente analizados por Dream House Finance, en donde se realizó un análisis exploratorio de éstos, incluido el tratamiento de los datos faltantes.

Procesados los datos se efectuó un análisis predictivo, en el cual se entrenaron los algoritmos Decision Tree y Naive Bayes en búsqueda del que tuviera mejor performance sobre el data set. Seguidamente se analizó la implementación del modelo en producción para probar el funcionamiento del algoritmo en productivo. Por último se observó como sería la implementación del mismo, donde se definieron los sectores de la empresa que intervendrán en el proceso y los roles que deberán cumplir para lograr el objetivo.

Apartados

Apartado 1. Gestión de datos en contextos organizacionales

1.1 Descripción de la organización

El presente trabajo se basó en la empresa Dream Housing Finance, que se desempeña en el rubro de préstamos hipotecarios dentro del ámbito privado en Estados Unidos desde 1992. Entre sus opciones, ofrece préstamos para adquirir una vivienda nueva o para refaccionar su vivienda actual.
Un préstamo hipotecario es un sistema de financiación que permite a una persona física adquirir un determinado bien o servicio financiando su costo a medio y largo plazo. Con dicha operación financiera en la cual Dream Housing Finance, el prestamista, entrega cierta cantidad de dinero a la otra parte, denominada prestatario, el cual hipoteca su propiedad a favor del prestamista y se compromete a devolver el capital prestado más intereses en los plazos y condiciones pactadas de ante mano.

1.2 Gestión de datos por parte de la organización

En el proceso de otorgamiento de préstamos hipotecarios intervienen distintas áreas de la empresa. El área comercial, la de operaciones, la de riesgo y la legal forman parte del circuito total.
El área comercial, es la fuerza de ventas y comercialización de los productos y servicios de acuerdo a las necesidades del cliente. Se encarga de captar nuevos usuarios, convertirlos en clientes y seguir manteniendo la relación con ellos. Son los responsables de recepcionar y validar toda la documentación solicitada para la operación a realizar. Fotocopia del documento de identidad, número de credencial de seguridad social, constancia salarial y último recibo de pago, resumen de cuentas bancarias y/o estados de cuentas y la solicitud debidamente completada y firmada. Toda esta información se pide tanto al solicitante como al coaplicante en caso de existir.
El área de operaciones recibe el legajo y se encarga de validar la información proporcionada por el cliente a través de llamadas telefónicas y búsqueda en internet. Verifica de esta forma los datos básicos, laborales, referencias económicas etc.
La carpeta pasa al área de riesgo, que evalúa y precalifica el préstamo. Puede aprobar o rechazar la solicitud según la documentación presentada para la evaluación. Además, este sector se encarga de valuar la tasa y los intereses de cada préstamo según el riesgo que consideren en cada caso.
Si la solicitud pasa el filtro de riesgo, el área legal se encarga de la validación, confección y seguimiento de la documentación correspondiente.
Una vez aprobado el préstamo y firmados los papeles correspondientes, el área de operaciones se encarga de la custodia de toda la documentación referente al préstamo, tales como constitución de garantía, documentación tanto del coaplicante como del titular del préstamo y los respectivos registros contables que implica el desembolso del préstamo.

1.3 Problemática de la organización y la gestión de los datos

Tal como se mencionó previamente el proceso inicia cuando el cliente entra a la sucursal y es atendido por un comercial en donde en primera instancia y según su intuición, conocimiento y sesgos, decide si se rechaza la solicitud o si se acepta inicialmente para continuar con el proceso. Tal como mencionan Render, Stair y Hanna (2014) la vida sería más sencilla si supiéramos sin lugar a dudas qué va a ocurrir en el futuro. El resultado de cualquier decisión dependería tan sólo de qué tan lógica y racional fuera la decisión, si de antemano supiéramos si el cliente va a pagar el préstamo o en los costos que tendrá que incurrir la organización para poder cobrar, ya no habría una decisión que tomar, sino que habría un único camino por delante siendo la vida más sencilla. Dado que no tenemos certeza de lo que va a ocurrir en el futuro, el resultado de cualquier decisión que tomemos dependerá de qué tan racional y lógica sea (Render, Stair y Hanna, 2014). Esta decisión de otorgar o no un préstamo tiene un costo no sólo administrativo, sino la posibilidad de otorgar un préstamo cuando no se debía o, a la inversa, que se debería haber otorgado el mismo, pero no se adjudicó. De lo que tenemos certeza es que la empresa otorga préstamos hipotecarios realiza un estudio detallado de la información que le solicita a los aplicantes, tales como si el solicitante trabaja en relación de dependencia o es independiente, niveles de ingresos propios y del coaplicante de existir, historial crediticio, valor del préstamo solicitado, estado civil, y todos los datos relacionados con el inmueble a hipotecar, como ser tipo de propiedad, tasación y/o ubicación. En caso de que se pase esa instancia se hace las derivaciones correspondientes para la validación del préstamo y de los intereses que se aplicarían en cada préstamo.

Dado que el proceso es lento y la información del solicitante debe pasar por varias áreas, la empresa se encuentra en la búsqueda de agilizar los procesos y lograr reducir costos. Dado estos motivos la organización se encuentra en búsqueda de pasar a un proceso de pre-validación de préstamos de manera virtual. En donde el cliente en vez de ir a una sucursal y llevar toda la documentación, que un comercial la reciba y haga un primer análisis donde la decisión se basa en instintos y su experiencia, pueda en unos segundos saber si está calificado para la obtención de un préstamo.
Con este proyecto la empresa lograría disminuir el tiempo de análisis del comercial en la primera instancia del proceso ya que los que asisten ya fueron aprobados y sólo tendrían que dedicarse a realizar la recepción de la documentación y sólo podrían rechazar si ven que alguno de los documentos presentados no cumple con los requisitos necesarios. Por otra parte, esto logrará que el volumen de casos que tiene que analizar el resto de las áreas disminuirá, por lo que podrán lograr dar un mejor análisis de cada caso y lograr disminuir tiempo para lograr que el cliente tenga una respuesta más ágil.
Pero para lograr esto se deberá utilizar de manera inteligente los datos. Es decir, dejar la toma de decisiones según la intuición y aplicar el análisis de los datos. Para esto la empresa deberá incurrir en el manejo de Big Data. En primera instancia Big Data es un conjunto de grandes volúmenes de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados (Laney, 2001). Pero, por otra parte, el Economista Diebld la define de la siguiente manera El fenómeno Big Data, es un crecimiento explosivo en el volumen, la velocidad y la variedad de datos y está en el corazón de la ciencia moderna. De hecho, la necesidad de lidiar con Big Data y la conveniencia de desbloquear la información oculta en su interior es ahora un tema clave en todas las ciencias,

posiblemente el tema científico clave de nuestro tiempo (Francis X. Diebold, 2020). Tal como menciona Diebold Big Data es un tema clave de nuestro tiempo, pero tiene sus ventajas y desventajas. El profesor universitario e historiador británico Cannadine (2020) menciona que los datos no son de todo importante, es lo que se hace con ellos lo que cuenta. Con la evolución del Big Data llegaron diversas formas de analizar os nuevos conjuntos de datos a los que ahora tenemos acceso. Como resultado, Big Data ha sido aclamado por su potencial para mejorar la toma de decisiones en campos desde los negocios hasta la medicina, permitiendo que los juicios y evaluaciones se basen cada vez más en información y análisis en lugar de intuición y conocimiento.

Por otro lado, Cannadine (2020) también comenta sobre otro punto de vista del Big Data, tal como mencionó en el pasado Sir Francis Bacon el conocimiento es poder, pero quizás el equivalente moderno es los datos son poder. Hoy el termino vigilancia de datos muestra cómo el modelo de arte de gobernar está cambiando en la era del Big Data. Hoy en día, la vigilancia rastrea a las personas a través de sus datos, y hay una carrera por los datos de la misma manera que alguna vez hubo una carrera por el petróleo. El propio Orwell, en cambio, esbozó un escenario más escalofriante en su novela clásica, 1984, publicada en 1949: Siempre ojos mirándote y la voz que te envuelve. Dormido o despierto, adentro o afuera, en el baño o en la cama, no hay escapatoria. Nada era tuyo excepto los unos centímetros cúbicos en tu cráneo.

El Big Data es un hecho y vino para quedarse, habrá organizaciones que deberán adaptarse más que otras, las empresas de Internet tienen las ventajas adecuadas de los datos para ingresar a la industria financiera. Frente a los desafíos, los bancos ya no pueden apegarse a las convenciones, por el contrario, deben establecer una estrategia de big data que jugará un papel importante en el proceso dinámico de toma de decisiones del riesgo crediticio de microempresas. Dado que proporciona respuestas a muchas preguntas que las empresas ni siquiera sabía que tenía. En otras palabras, proporciona un punto de referencia. Con una cantidad tan grande de información, los datos pueden ser moldeados o probados de cualquier manera que la empresa considere adecuada. Al hacerlo, las organizaciones son capaces de identificar los problemas de una forma más comprensible.

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades. Eso, a su vez, conduce a movimientos de negocios más inteligentes, operaciones más eficientes, mayores ganancias y clientes más felices. Las empresas con más éxito con Big Data consiguen valor de las siguientes formas reducción de coste. Las grandes tecnologías de datos, como Hadoop y el análisis basado en la nube, aportan importantes ventajas en términos de costes cuando se trata de almacenar grandes cantidades de datos, además de identificar maneras más eficientes de hacer negocios.

Por otra parte ayudara a tomar mejores decisiones de manera más rápida. Con la velocidad de Hadoop y la analítica en memoria, combinada con la capacidad de analizar nuevas fuentes de datos, las empresas pueden analizar la información inmediatamente y tomar decisiones basadas en lo que han aprendido.

Desde el sistema de indicadores de datos financieros estáticos hasta el sistema de indicadores dinámicos de Internet, la extracción de información relacionada con la calificación crediticia de los clientes, los bancos pueden hacer que los riesgos crediticios de crédito de microempresas sean lo más bajos posible. En comparación con el modo tradicional de toma de decisiones sobre el riesgo de crédito, este modo es mucho más barato, más fácil y más preciso (Junxuan Zhu & Zhe Huang, 2014).

Pero para lograr este cambio de paradigma la empresa deberá afrontar diversos retos y obstáculos en la utilización del Big Data entre ellos se encuentran: la calidad de los datos y el tratamiento de los datos faltantes, el conocimiento para procesar y analizar los datos, las muestras obtenidas no siempre representan a la población que se busca analizar, los datos deben ser interdependientes y no tener el mismo grado de dependencia, la velocidad del procesamiento de los datos, la capacidad de adaptar los procesos de producción de estadísticas a las nuevas fuentes de información y que toda la organización esté dispuesta a los cambios.

Por estos motivos, la generación y el análisis de datos aumentan su relevancia en el mundo en que vivimos. La empresa Dream Housing Finance tiene que empezar a utilizar técnicas de Big Data y Data mining para procesar y analizar información valiosa en datos estructurados y no estructurados. Valerse de éstas técnicas en forma efectiva impulsará la innovación en el modelo comercial y el modelo de toma de decisiones.

Por otra parte, Dream Housing Finance no sólo tendrá que incurrir en Big Data sino que también tendrá que incurrir en la gobernanza de datos. El gobierno de datos es la gestión global de la disponibilidad, relevancia, usabilidad, integridad y seguridad de los datos en una empresa. Ayuda a las organizaciones a gestionar sus conocimientos de información y a responder preguntas, tales como ¿Qué sabemos sobre nuestra información?, ¿De dónde provienen estos datos? Y ¿Se adhieren estos datos a las políticas y reglas de la empresa?.

El gobierno de datos proporciona un enfoque holístico para administrar, mejorar y aprovechar la información de forma que pueda ayudarnos a ganar percepción y generar confianza en decisiones y operaciones empresariales.

Lograr una buena gobernabilidad y gestión de datos la empresa deberá abordar la gestión de los datos como lo que son en realidad, un activo de gran valor tanto a nivel operativo como para crear valor de mercado y convertirlos en una información crítica para el negocio. Esa óptima gestión de datos clave para el éxito empresarial requiere de un marco que acoja un gobierno de datos, entendido como el ejercicio de diseñar, controlar y monitorizar todo lo relativo a los datos desde un enfoque holístico, en el que participen los implicados, desde el gobierno corporativo de la empresa y el departamento de IT hasta un consejo de gestión de datos que represente a las partes interesadas.

La función de gobierno de datos es conseguir que todas las funciones de datos se realicen del modo más eficiente, cumpliendo con lo planeado. Se trata, de asegurar que los datos cumplen con las demandas, al tiempo que se consigue una reducción de costes en lo que respecta a su gestión y a su protección, éste último un aspecto importante en lo que respecta al cumplimiento de normativas y a la preservación de la privacidad.

El gobierno de datos es necesario ya que permite asegurar la integridad. Además, evita y previene incoherencias entre distintos sistemas o aplicaciones, con la ventaja que, por ejemplo, ello supone para que no falten datos a la hora de operar, de hacer evaluaciones o de ofrecer un determinado servicio o información.

Ayuda a responder a las demandas actuales. Establecer un marco para la gobernanza de datos nos ayuda a conseguir una mayor disponibilidad, facilidad de uso, consistencia, integridad y seguridad de los datos, requisitos clave para apoyar las iniciativas más actuales de BI, que normalmente requieren aplicaciones rápidas, con un acceso en tiempo real a los datos.

Permite agregar valor. Un plan de data governance, por último, ayuda a definir y establecer los diferentes tipos de comunicación necesaria para agregar valor a la organización a partir de una visión global capaz de transformar el negocio en su conjunto. Los equipos de

gestión podrán tomar decisiones informadas basadas en datos más fiables. No olvidemos que la información crítica es relevante, si no esencial, para tomar decisiones.
Cuando el gobierno de datos es deficiente o simplemente se carece de él, los datos no se integran en un concepto holístico del conocimiento de la información y su control, que entonces se realiza por departamentos o por sistemas, se convierte en una tarea pendiente. Por lo tanto, se pierde ese enfoque o visión general, esenciales para lograr una necesaria coherencia.
Dentro de este contexto, ignorar el carácter decisivo del gobierno de datos es una vía más directa hacia el descontrol en la gestión de los datos. Por el contrario, el data governance (el corazón de la gestión datos), cumple una función de control y coordinación interactiva entre las distintas áreas de la empresa, definiendo roles y responsabilidades, y estableciendo estándares, políticas y procesos de forma consensuada.

Apartado 2. Descripción metodológica

2.1. Recopilación de la información

El objetivo del presente trabajo es predecir si se otorgará o no un préstamo hipotecario en la empresa Dream House Finance en forma virtual. Para el estudio se utilizará una base de datos obtenida de Kaggle, en la cual se emplearán diversas técnicas de data mining o Minería de Datos que, en términos simples, consiste en encontrar patrones útiles en los datos. Es una palabra de moda, existe una amplia variedad de definiciones y criterios para la minería de datos. La minería de datos también se conoce como descubrimiento de conocimiento, aprendizaje automático y analítica predictiva (Bala Deshpande y Vijay Kotu,2014, p.2). Por lo que se podría decir que la minería de datos se encarga de resolver problemas mediante el análisis de datos ya presentes, es decir de una base de datos.
En este trabajo se utilizó el dataset Bank_Loan, obtenido de Kagle publicado el 2019-11-22. Esta base de datos contiene los datos de 614 solicitantes de préstamos hipotecarios de la empresa Dream House Finance. ver Ilustración 1. Dicho dataset está compuesto por las solicitudes de prestamos hipotecarios ya analizados por la empresa, los mismos cuentan con la información presentada de los solicitantes y la información que recauda la empresa sobre los mimos, entre ellos se encuentran ID del préstamo, genero del solicitante, si se encuentra casado, si tiene dependientes, el nivel educativo, si trabaja en relación de dependencia, los ingresos del solicitante, los ingresos del coaplicante, monto del préstamo, duración del préstamo, historial crediticio, área geográfica de la propiedad y si se otorgo e préstamo.

Ilustración 1- Variables del Data Set

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Loan_ID            614 non-null    object
 1   Gender             601 non-null    object
 2   Married            611 non-null    object
 3   Dependents         599 non-null    object
 4   Education          614 non-null    object
 5   Self_Employed      582 non-null    object
 6   ApplicantIncome    614 non-null    int64
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object
 12  Loan_Status        614 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

Fuente: elaboración propia

La misma cuenta con datos mixtos, cuantitativos (valor del préstamo, ingresos del solicitante, etc.) y cualitativos (si es casado, es trabajador en relación de dependencia, etc.). Por otra parte, la base de datos, en formato CSV, cuenta con un total de 7881 datos y 13 campos, teniendo datos faltantes en algunos de estos, ver Ilustración 2.

Ilustración 2 – Análisis de la base de datos

```
In [5]: details = summary(df)
        details
```

```
Data shape: (614, 13)
```

Out[5]:

	Type	Total count	Null Values	Distinct Values	Missing Ratio	Unique Values	Skewness	Kurtosis
Loan_ID	object	614	0	614	0.000000	[[LP001002, LP001003, LP001005, LP001006, LP00...	NaN	NaN
Gender	object	601	13	3	2.117264	[[Male, Female, nan]]	NaN	NaN
Married	object	611	3	3	0.488599	[[No, Yes, nan]]	NaN	NaN
Dependents	object	599	15	5	2.442997	[[0, 1, 2, 3+, nan]]	NaN	NaN
Education	object	614	0	2	0.000000	[[Graduate, Not Graduate]]	NaN	NaN
Self_Employed	object	582	32	3	5.211726	[[No, Yes, nan]]	NaN	NaN
ApplicantIncome	int64	614	0	505	0.000000	[[5849, 4583, 3000, 2583, 6000, 5417, 2333, 30...	6.539513	60.540676
CoapplicantIncome	float64	614	0	287	0.000000	[[0.0, 1508.0, 2358.0, 4196.0, 1516.0, 2504.0,...	7.491531	84.956384
LoanAmount	float64	592	22	204	3.583062	[[nan, 128.0, 66.0, 120.0, 141.0, 267.0, 95.0,...	2.677552	10.401533
Loan_Amount_Term	float64	600	14	11	2.280130	[[360.0, 120.0, 240.0, nan, 180.0, 60.0, 300.0...	-2.362414	6.673474
Credit_History	float64	564	50	3	8.143322	[[1.0, 0.0, nan]]	-1.882361	1.548763
Property_Area	object	614	0	3	0.000000	[[Urban, Rural, Semiurban]]	NaN	NaN
Loan_Status	object	614	0	2	0.000000	[[Y, N]]	NaN	NaN

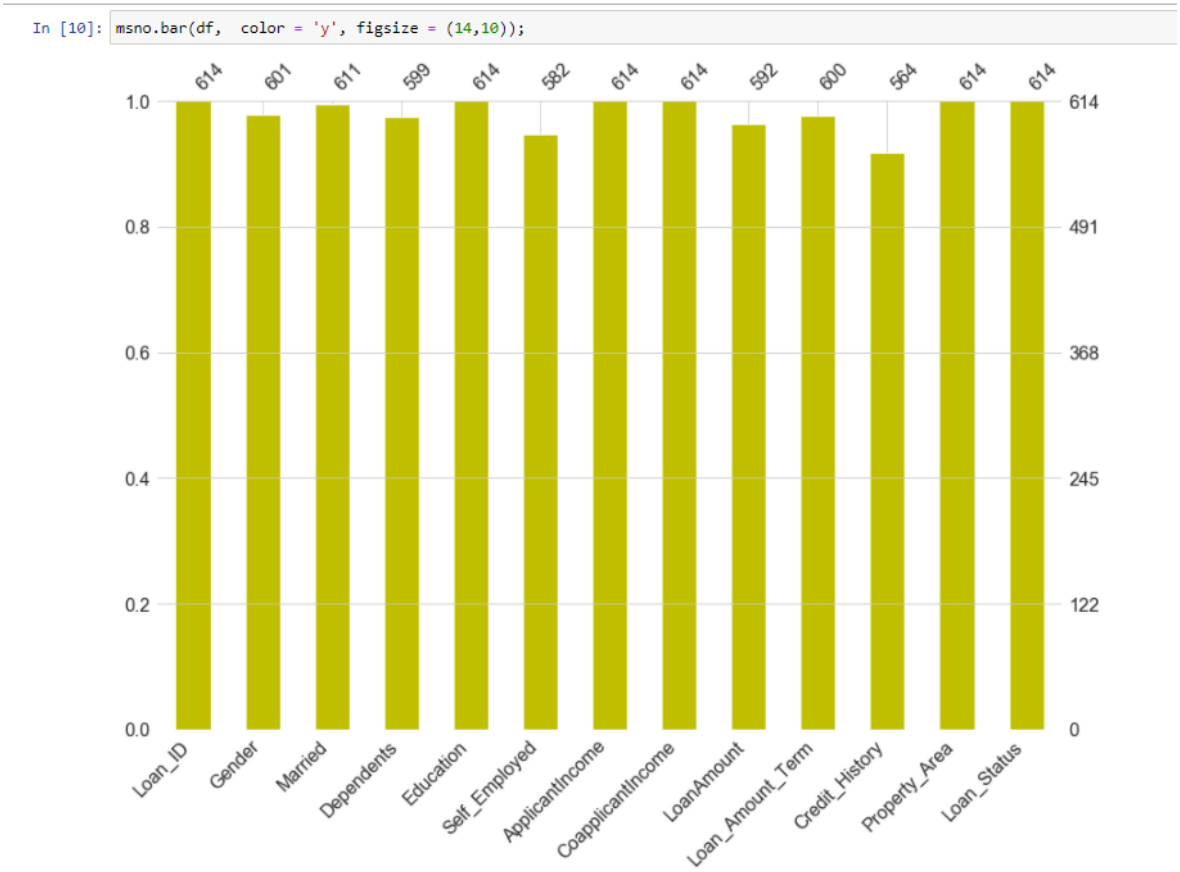
Fuente: elaboración propia

2.2. Procesamiento de la información

Para el análisis del dataset se utilizarán todos los campos que contiene la base, Loan Id; Género; Estado Civil; Dependientes; Educación; Self Emplyed; Ingresos del Aplicante; Ingresos del Coaplicante; Valor del préstamo (en miles); Duración del Préstamo; Historial Crediticio; Locación geográfica del Área; Estado del Préstamo.
Para el procesamiento de la base de datos se utilizó el Software Azure Machine Learning, Jupiter Notebook 6.0.3, Power BI Desktop y Excel del paquete office 365 de Microsoft.

La base de datos contaba con 149 datos perdidos, distribuidos de la siguiente manera: Gender 13 missing; Maried 3; Depends 15; Sefl_Employe 32; Loan_Amount 22; Loan_Amount_Term; Credit_History 50, ver Ilustración 3.

Ilustración 3- Visualización de los datos faltantes



Fuente: elaboración propia

Lo primero que se trabajó fue para todos los atributos se creó una columna llamada nombre del atributo _Missing en donde rellenó con la función =SI con 1 y 0, siendo 0 indicador de que el campo estaba vacío.
El tratamiento que se les dio fue el siguiente, a todas las columnas tenían o no valores perdidos se le creó al lado una columna llamada nombre del atributo _Missing en donde rellenó con la función =SI con 1 y 0, siendo 0 indicador de que el campo estaba vacío.
Una vez que ya se tenía todas las columnas indicando que valor estaba faltante, se prosiguió a rellenar los campos vacíos en Jupiter Notebook. Para las variables Gender, Married, Dependents, Self_Employed, Credit_History y Loan_Amount_Term se rellenaron con la moda de cada variable, de esta forma si el campo estaba vacío traía el valor con mayor frecuencia. En cambio, para el atributo Loan_Amount se aplicó la MEDIANA para que los campos que tuvieran missing values los completará con el valor de la variable de posición central.

2.3. Análisis de la información

Tal como se mencionó previamente la base de datos fue procesada quedando compuesta por 24 campos, incluidos el id y el label, cada uno con 614 datos y sin campos vacíos, como se muestra en la Ilustración 4.

Ilustración 4 - Composición de la base de datos procesada

```
In [45]: details = summary(df)
         details

         Data shape: (614, 24)
```

Out[45]:

	Type	Total count	Null Values	Distinct Values	Missing Ratio	Unique Values	Skewness	Kurtosis
Loan_ID	object	614	0	614	0.0	[[LP001002, LP001003, LP001005, LP001006, LP00...	NaN	NaN
Gender	object	614	0	2	0.0	[[Male, Female]]	NaN	NaN
Gender_missing	int64	614	0	2	0.0	[[1, 0]]	-6.668550	42.608339
Married	object	614	0	2	0.0	[[No, Yes]]	NaN	NaN
Married_Missing	int64	614	0	2	0.0	[[1, 0]]	-14.235914	201.316988
Dependents	float64	614	0	4	0.0	[[0.0, 1.0, 2.0, 3.0]]	1.015551	-0.347376
Dependents_missing	int64	614	0	2	0.0	[[1, 0]]	-6.176135	36.262758
Education	object	614	0	2	0.0	[[Graduate, Not Graduate]]	NaN	NaN
Education_Missing	int64	614	0	1	0.0	[[1]]	0.000000	0.000000
Self_Employed	object	614	0	2	0.0	[[No, Yes]]	NaN	NaN
Self_Employed_Missing	int64	614	0	2	0.0	[[1, 0]]	-4.040073	14.368984
ApplicantIncome	int64	614	0	505	0.0	[[5849, 4583, 3000, 2583, 6000, 5417, 2333, 30...	6.539513	60.540676
ApplicantIncome_Missing	int64	614	0	1	0.0	[[1]]	0.000000	0.000000
CoapplicantIncome	int64	614	0	287	0.0	[[0, 1508, 2358, 4196, 1516, 2504, 1526, 10968...	23.917976	582.666878
CoapplicantIncome_Missing	int64	614	0	1	0.0	[[1]]	0.000000	0.000000
LoanAmount	float64	614	0	204	0.0	[[146.41216216216216, 128.0, 66.0, 120.0, 141....	2.726601	10.896456
LoanAmount_Missing	int64	614	0	2	0.0	[[0, 1]]	-5.006862	23.144049
Loan_Amount_Term	float64	614	0	10	0.0	[[360.0, 120.0, 240.0, 180.0, 60.0, 300.0, 480...	-2.402112	6.924993
Loan_Amount_Term_Missing	int64	614	0	2	0.0	[[1, 0]]	-6.409453	39.208795
Credit_History	float64	614	0	2	0.0	[[1.0, 0.0]]	-2.021971	2.095179
Credit_History_Missing	int64	614	0	2	0.0	[[1, 0]]	-3.068326	7.438848
Property_Area	object	614	0	3	0.0	[[Urban, Rural, Semiurban]]	NaN	NaN
Property_Area_Missing	int64	614	0	1	0.0	[[1]]	0.000000	0.000000
Loan_Status	object	614	0	2	0.0	[[Y, N]]	NaN	NaN

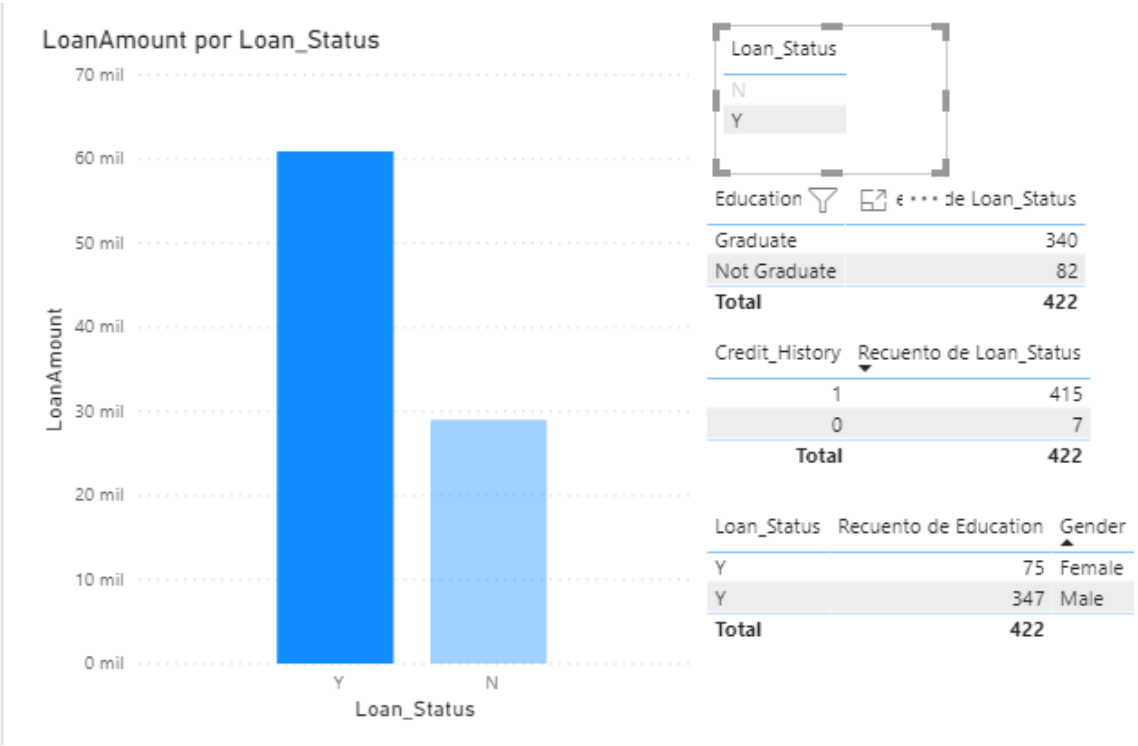
Fuente: elaboración propia

Luego del procesamiento de la base de datos se realizó un análisis de las variables categóricas Gender, Maried, Education, Self_Empoyed, Credit histori y de Property_Area,
La variable Gender,la cual originalmente tenía 601 valores de los cuales tenía 13 faltantes paso a tener la siguiente distribución: 502 hombres y 112 mujeres, siendo un 82% de los solicitantes a préstamos son de género masculinos y el resto, el 18%, femeninos.
En cuanto a la variable Maried que originalmente tenía 611 valores y 3 campos vacios paso a estar compuesta por 401 individuos que están casados y 213 que no. Por lo que la distribución de la variable quedo conformada por un 65% por los solicitantes que se encuentran casados y un 35% que no.
En el caso de la variable Education no se le realizó ningún proceso ya que esta variable no tenía datos faltantes. La misma quedó compuesta las variables yes y no haciendo referencia a si están graduado o no. La distribución de la misma quedo compuesta por 78% de los solicitantes están graduados y un 22% no.
En cuanto a la variable Self_Employed, originalmente poseía 32 datos faltantes de los 614 campos. Luego del procesamiento la variable paso a tener 532 personas que no trabajan por cuenta propia y 82 que sí, teniendo un 87% de autónomos y un 13% de trabajadores en relación de dependencia.
Para la variable Credit_History, en la cual se clasifica si el historial crediticio cumple con las pautas siendo 1 si y 0 no, originalmente tubo 50 valores faltantes, luego del

procesamiento pasó a estar compuesta de la siguiente manera 525 si y 89 no, siendo un 86% para los casos afirmativos y un 14% para los no.
Por último la variable categorica Property_Area la cual se clasifica en 3 tipos, Urbana, Rural y Semiurbana, tampoco tubo casos de valores faltantes por lo que no se le realizó un tratamiento. Por tales motivos la distribución de las variables quedó compuesta por 233 en área semiurbana, 202 en urbanas y 179 en área rural. Siendo un 38% los casoso que se solicita el préstamo para una vivienda en el área Semiurbana, un 33% para los casos en áreas Urbanas y un 29% para el área Rural.

En el caso de la variable Loan_Status, la cual es la variable que dice si se otorgó el préstamo y nuestra variable a predecir, tiene una distribución de 422 casos donde si se otorgó y 192 en los que no. Por otra parte entre los 422 casos se puede observar que 340 personas tienen estudios y 82 no, 415 vidas tenían un buen historial crediticio y 7 no, y del total de los préstamos otorgados 347 fueron a hombres y solo 75 a mujeres, esto se puede observar en la Ilustración 5.

Ilustración 5 - Relación de las variables Loan_Status, Education, Credit_History y Gender



Fuente: elaboración propia

Apartado 3. Implementación de modelos de aprendizaje automático

3.1. Puesta en producción de un modelo

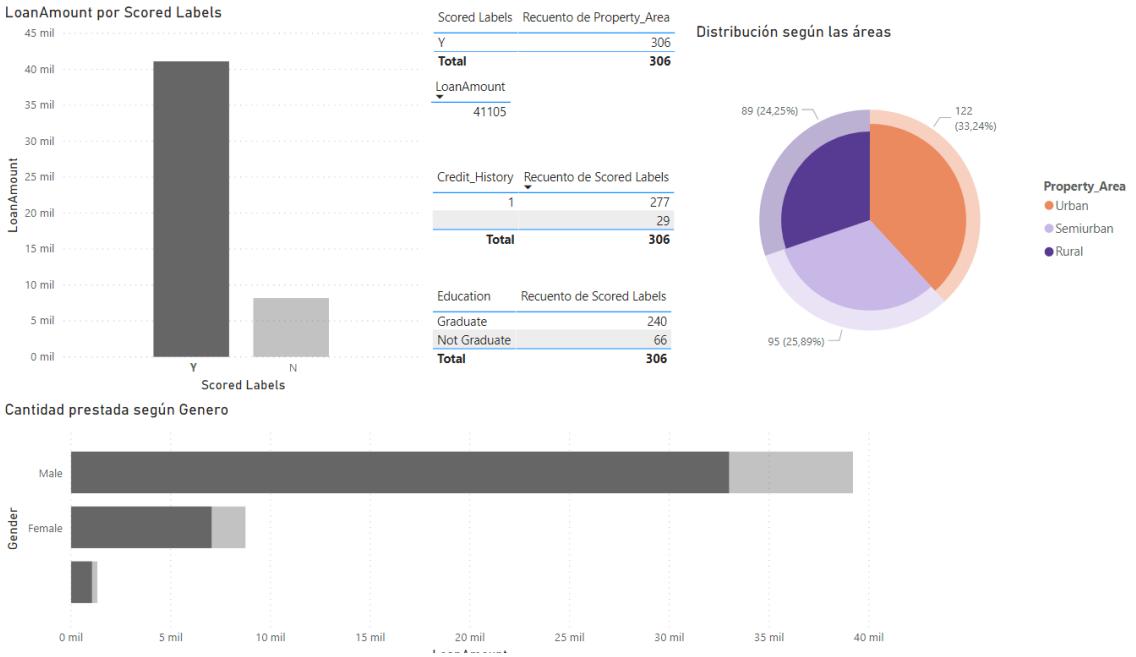
Una vez finalizada la etapa de procesamiento de la información y su análisis posterior se prosiguió con el entrenamiento de los algoritmos en Microsoft Azure Machine Learning Studio. Se utilizaron 2 algoritmos de aprendizaje automático para poder seleccionar cual es mejor en relación con la exactitud (accurracy) de los mismos. Los algoritmos de clasificación seleccionados fueron Decision Tree y Naive Bayes.
Decision Tree: Los árboles de decisión abordan el problema de clasificación dividiendo datos en subconjuntos "más puros" basados en los valores de los atributos de entrada. Los atributos que ayudan a lograr los niveles más limpios de dicha separación son consideradas significativas en su influencia sobre la variable objetivo y el final en la raíz y los niveles más cercanos a la raíz del árbol. El modelo de salida es un marco de árbol que se puede utilizar para la predicción de nuevos no etiquetados datos.
Naive Bayes: Los algoritmos bayesianos ingenuos proporcionan una forma probabilística de construir un modelo. Este enfoque calcula la probabilidad para cada valor de la variable de clase para valores dados de variables de entrada. Con la ayuda de probabilidades condicionales, para un registro desconocido dado, el modelo calcula el resultado de todos los valores de las clases de destino y obtiene un ganador previsto.
La evaluación de los resultados que se obtuvieron fueron por medio de la comparación del Accuracy de los modelos ya mencionados. Lo primero que se realizó fue ejecutar desde Microsoft Azure Machine Learning Studio la base de datos procesada, utilizando un Split data con 80% de los datos para entrenamiento y un 20 % para entrenamiento y usando una semilla 2001. Además, se usaron los siguientes parámetros, Decision Tree se utilizaron 50 árboles y con una profundidad de 150. En el caso de Naive Bayes se utilizaron 2 iteraciones. Como resultado del entrenamiento de los algoritmos se obtuvieron los siguientes resultados. Para Decision Tree el resultado fue de 0.770 mientras que Naive Bayes fue de 0.6890. Dado que Decision Tree tuvo el mejor performance se guardó dicho modelo para poder usar lo en la implementación.
Para la prueba de la implementación se utilizó otra base de datos de Kaggle de la misma competencia, la cual tiene como finalidad la de probar el modelo. La misma cuenta con 367 casos para analizar el otorgamiento de préstamos. A dicha base se le realizó un procedimiento similar al data set de pruebas, desde Excel se le agregaron las columnas de Missing_, ya que esta base debe tener las mismas variables que la base que se usó para entrenar el modelo. Una vez que ya se proceso la nueva base de datos, la misma se tuvo que subir a Azure, en formato CSV, y cargar el algoritmo entrenado para procesar la.
Dado que estos valores no son reales y la competencia de Kaggle se encuentra cerrada, lo cual no permite subir los resultados obtenidos del modelo para ver como fue la performance del mismo, solo se puede analizar los resultados obtenido.

3.2. Visualización de Insights

Entre los resultados obtenidos de haber procesado la nueva base de datos se puede observar que de los 367 casos según el modelo entrenado se otorgarían 306 préstamos, ver Ilustración 6. De los cuales 277 casos cumplen con los requisitos de poseer un buen historial crediticio y 29 casos en donde ese campo se encurta vacío. Por otra parte, se

puede observar como la mayor parte de los solicitantes de los prestamos son para inmuebles en zonas urbanas, siendo un 38.15% del total y un 33.24% para el caso de los pre aprobados, mientras que para las áreas de semiurbanas 30.25% del total y 24.25% para el caso de los pre aprobados, y en el caso del área rural 31.61% del total y un 25.89% para el caso de los pre aprobados. Además, otro dato que resalta al analizar los resultados es que más del 50% de las solicitudes son del género masculino.

Ilustración 6 - Resultados del modelo entrenado



Fuente: elaboración propia

3.3. Metodologías ágiles

La empresa Dream Housing Finance posee un proceso de otorgamiento de préstamos hipotecarios lento y con mucho volumen de trabajo, en donde la información del solicitante debe ser analizada en primera instancia por un comercial y ver si cumple con los requisitos iniciales y luego pasar por el resto de las áreas. Dado estos motivos la financiera se encuentra en la búsqueda de agilizar los procesos y lograr reducir costos. Parte de este cambio consiste en contar con un proceso de pre-validación de préstamos en forma virtual. Aquí el postulante, en vez de dirigirse personalmente a la sucursal con toda la documentación para que un comercial la reciba y realice un primer análisis donde la decisión se basa en instintos y experiencia, ingresa una mínima cantidad de datos en una plataforma virtual y en segundos sabe si está calificado. De esta manera se podrán descartar automáticamente algunos casos y profundizar con la información y el análisis de los casos pre-validados.

Para lograr la implementación de la pre-validación de los préstamos va a ser necesario incurrir en metodologías ágiles. Considero que la más adecuada para este caso es la Extreme Programming (XP). Esta metodología es la conveniente porque se centra en potenciar las relaciones interpersonales como clave para el éxito en desarrollo de software, promoviendo el trabajo en equipo, preocupándose por el aprendizaje de los

desarrolladores, y propiciando un buen clima de trabajo. Dicha metodología ágil se basa en realimentación continua entre el cliente y el equipo de desarrollo, comunicación fluida entre todos los participantes, simplicidad en las soluciones implementadas y coraje para enfrentar los cambios.

Para este proyecto estarán presentes las áreas de IT, comercial, operaciones, riesgo y mejora continua, las cuales tendrán que cumplir los roles de programador, cliente, tester, tracker, coach y líder del proyecto.

El cliente interno, el sector comercial, operaciones y riesgo, transmite sus requerimientos y necesidades al área de mejora continua quien baja a detalle el pedido, lo divide en iteraciones, comienza a trabajar en conjunto con el resto de los involucrados en el proyecto y realizará su seguimiento durante todo el ciclo de vida del mismo. Mejora continua será en resumen el Project leader, coach y tracker de acuerdo al momento y circunstancias. IT definirá y se encargará del hardware necesario, del software licenciado, de producir el código de programación, implementar lo producido y realizar todos los cambios solicitados tantas veces como sea necesario hasta lograr que el producto terminado cumpla con las expectativas del cliente interno. Los usuarios de comercial y riesgo serán los encargados de armar los sets de datos para las pruebas y realizarán las mismas. También serán los responsables de solicitar modificaciones para aportar mayor valor al negocio y/o detectar fallas durante las pruebas. Su nexo con IT será mejora continua. Cuando todos los testeos pasen el filtro del cliente interno, los programas se implementarán y se difundirán los resultados de las pruebas realizadas.

Mejora continua en su función de tracker tendrá que proporcionar realimentación al equipo. Verifica el grado de acierto entre las estimaciones realizadas y el tiempo real dedicado, para mejorar futuras estimaciones. Deberá realizar el seguimiento del progreso de cada iteración. Además, como coach deberá proveer guías al equipo de forma que se apliquen las prácticas XP y se siga el proceso correctamente ayudando a que el equipo trabaje efectivamente creando las condiciones adecuadas. Su labor esencial es de coordinación.

Al igual que otras metodologías de gestión de proyectos, tanto Ágiles como tradicionales, el ciclo XP incluye seis etapas, entre las cuales se encuentran la fase de exploración, planificación de las entregas, iteraciones, producción, mantenimiento y muerte del proyecto. La fase de exploración, los clientes internos plantean a grandes rasgos las historias de usuario que son de interés para la primera entrega del producto. Al mismo tiempo el equipo de desarrollo se familiariza con las herramientas, tecnologías y prácticas que se utilizarán en el proyecto. Se prueba la tecnología y se exploran las posibilidades de la arquitectura del sistema construyendo un prototipo. La fase de exploración toma de pocas semanas a pocos meses, dependiendo del tamaño y familiaridad que tengan los programadores con la tecnología.

En la fase de planificación de la entrega, el cliente interno establece la prioridad de cada historia de usuario, y correspondientemente, los programadores realizan una estimación del esfuerzo necesario de cada una de ellas. Se toman acuerdos sobre el contenido de la primera entrega y se determina un cronograma en conjunto con el cliente. Una entrega debería obtenerse en no más de tres meses. Esta fase dura unos pocos días. Las estimaciones de esfuerzo asociado a la implementación de las historias la establecen los programadores utilizando como medida el punto. Un punto, equivale a una semana ideal de programación. Las historias generalmente valen de 1 a 3 puntos. Por otra parte, el equipo de desarrollo mantiene un registro de la “velocidad” de desarrollo, establecida en puntos

por iteración, basándose principalmente en la suma de puntos correspondientes a las historias de usuario que fueron terminadas en la última iteración. La planificación se puede realizar basándose en el tiempo o el alcance. La velocidad del proyecto es utilizada para establecer cuántas historias se pueden implementar antes de una fecha determinada o cuánto tiempo tomará implementar un conjunto de historias. Al planificar por tiempo, se multiplica el número de iteraciones por la velocidad del proyecto, determinándose cuántos puntos se pueden completar. Al planificar según alcance del sistema, se divide la suma de puntos de las historias de usuario seleccionadas entre la velocidad del proyecto, obteniendo el número de iteraciones necesarias para su implementación.
En la fase de iteraciones, incluye varias iteraciones sobre el sistema antes de ser entregado. El Plan de Entrega está compuesto por iteraciones de no más de tres semanas. En la primera iteración se puede intentar establecer una arquitectura del sistema que pueda ser utilizada durante el resto del proyecto. Esto se logra escogiendo las historias que fuercen la creación de esta arquitectura, sin embargo, esto no siempre es posible ya que es el cliente quien decide qué historias se implementarán en cada iteración (para maximizar el valor de negocio). Al final de la última iteración el sistema estará listo para entrar en producción. Los elementos que deben tomarse en cuenta durante la elaboración del Plan de la Iteración son: historias de usuario no abordadas, velocidad del proyecto, pruebas de aceptación no superadas en la iteración anterior y tareas no terminadas en la iteración anterior. Todo el trabajo de la iteración es expresado en tareas de programación, cada una de ellas es asignada a un programador como responsable, pero llevadas a cabo por parejas de programadores.

En la etapa de producción, requiere de pruebas adicionales y revisiones de rendimiento antes de que el sistema sea trasladado al entorno del cliente. Al mismo tiempo, se deben tomar decisiones sobre la inclusión de nuevas características a la versión actual, debido a cambios durante esta fase. Es posible que se rebaje el tiempo que toma cada iteración, de tres a una semana. Las ideas que han sido propuestas y las sugerencias son documentadas para su posterior implementación. Una vez que el proyecto esté en condiciones cumpliendo los requerimientos necesarios para llevar lo a implementar, es decir que tenga una cierta performance y sea de fácil uso para los usuarios, se deberá presentar el mismo a los directivos de la empresa para que lo aprueben. Una vez aprobado se pasará a implementar en productivo donde se evaluará si el mismo funciona adecuadamente con datos reales. En caso de que al momento de evaluar lo en producción los resultados no son acordes a los deseados se deberá volver atrás y evaluar que es lo que no está funcionando, pero en el caso de que el rendimiento del algoritmo en productivo sea satisfactorio solo quedará dar mantenimiento al mismo.
En la fase de mantenimiento, mientras la primera versión se encuentra en producción, el proyecto XP debe mantener el sistema en funcionamiento al mismo tiempo que desarrolla nuevas iteraciones. Para realizar esto se requiere de tareas de soporte para el cliente. De esta forma, la velocidad de desarrollo puede bajar después de la puesta del sistema en producción. La fase de mantenimiento puede requerir nuevo personal dentro del equipo y cambios en su estructura.
Como última fase del proyecto se encuentra la etapa de muerte del proyecto, la cual toma lugar cuando el cliente no tiene más historias para ser incluidas en el sistema. Esto requiere que se satisfagan las necesidades del cliente en otros aspectos como rendimiento y confiabilidad del sistema. Se genera la documentación final del sistema y no se realizan

más cambios en la arquitectura. La muerte del proyecto también ocurre cuando el sistema no genera los beneficios esperados por el cliente o cuando no hay presupuesto para mantenerlo.

Conclusiones

El objetivo del presente trabajo era intentar responder a la pregunta ¿Cuáles técnicas de minería de datos se podrán utilizar para realizar una mejor clasificación binaria de los solicitantes a préstamos hipotecarios para determinar si es posible su otorgamiento de manera virtual?
Para llegar a una conclusión primero se debió realizar un saneamiento de la base de datos en Excel, donde se completaron los campos vacíos y se crearon nuevas variables. Una vez finalizada la limpieza, se aplicaron diversas técnicas de minería de datos desde Azure Machine Learning Studio, entre ellas Decision Tree y Naive Bayes, para poder ver cuál de todas ellas se adapta mejor a esta situación. Dado que lo buscado era realizar una clasificación binaria, es decir clasificar entre sí o no, era cuestión de analizar cuál de los algoritmos utilizados tendría mejor performance. En este caso, luego de optimizar cada algoritmo, se obtuvo que el mejor Accuracy fue Decision Tree.
Si se decide utilizar este análisis para el otorgamiento de préstamos, sería interesante a futuro volver a plantearlo con el agregado de la evolución de los cobros de cuotas para detectar cuáles préstamos incurrieron en mora.

Referencias bibliográficas

- Breiman l. (2001). random forests. machine learning 45(1) 5–32.
- Peterson, L. (2009). K-Nearest Neighbors. Scholarpedia. Recuperado: http://www.scholarpedia.org/article/K-nearest_neighbor.
- Kotu, V., & Deshpande, B. (2014). Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann.
- Render,B., Stair, R. & Hanna M. E., (2014). Métodos Cuantitativos Para los Negocios.
- Pearson Pereira, S., Hernández Arteag, I., Zambrano, S., Hidalgo Troya, A. & Pérez, J., (2016)
- Descubrimiento de Patrones de Desempeño Académico, Universidad Cooperativa de Colombia Cannadine, D. (2020), "Big Data," Behind the Buzzwords , BBC Radio 4 Podcast, 4 August 2020, https://www.bbc.co.uk/programmes/m000lghg.
- Diebold, F.X. (2000), "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting," Eighth World Congress of the Econometric Society, Seattle, August. http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF.
- Diebold, F.X. (2003), "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Reichlin and Watson," In M. Dewatripont, L.P. Hansen and S. Turnovsky (eds.), Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society, Cambridge University Press, 115-122.
- Laney, D. (2001), "3-D Data Management: Controlling Data Volume, Velocity and Variety," META Group Research Note, February 6. http://goo.gl/Bo3GS.
- Junxuan Zhu1 & Zhe Huang1(2014), " Banks' Micro Enterprises Loan Credit Risk Decision-making Model Innovation in the Era of Big Data and Internet Finance", School of Management, Shanghai University of Engineering Science, Shanghai, China.
- Francis X. Diebold (2020), "On the Origin(s) of the Term "Big Data"", University of Pennsylvania, 2020
- Sato, D., Bassi, D., Bravo, M., Goldman, A., & Kon, F. Experiences tracking agile projects: an empirical study. Journal of the Brazilian Computer Society, 12(3), 45-64, 2006.
- Beck, K. Extreme Programming Explained: Embrace Change [1ª ed.]. Addison Wesley, Stoughton, 1999.
- Ronald, J. (2012). What is extreme programming [Internet], Disponible desde http://xprogramming.com/ what-is-extreme-programming/ [Acceso Junio 1, 2013].
- Beck, K., & Andres, C. Extreme Programming Explained: Embrace Change [2ª ed.]. Addison Wesley, Stoughton, 2004.
- L. Bottou, et al., Comparison of Classifier Methods: A Case Study in Handwritten Di it Recognition, Proceedings of the 12th International Conj?Terence on Pattern Recognition, 11, Jerusalem, Israel, Oct 9-13, 1994, 77-82.
- Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., &LichtendahlJr, K. C. (2018). Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons.